



Anhanguera

Medidas de
associação

Felipe
Figueiredo

Correlação

Medidas de associação

Correlação Linear

Felipe Figueiredo

UNIAN - Centro Universitário Anhanguera de Niterói

- 1 Correlação
 - Associação entre duas variáveis
 - Covariância entre duas amostras
 - Coeficiente de correlação de Pearson

- 1 Correlação
 - Associação entre duas variáveis
 - Covariância entre duas amostras
 - Coeficiente de correlação de Pearson

Tipos de variáveis envolvidas



Anhangüera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação

Covariância entre
duas amostras

Pearson

- Considere duas amostras X e Y , de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
 - X : variável independente (ou variável explanatória)
 - Y : variável dependente (ou variável resposta)

- ◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡

Tipos de variáveis envolvidas



Anhangüera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação

Covariância entre
duas amostras

Pearson

- Considere duas amostras X e Y , de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
 - X : variável independente (ou variável explanatória)
 - Y : variável dependente (ou variável resposta)

- Como definir (e mensurar!) o grau de associação entre duas variáveis aleatórias (VAs)?
- Se uma VA é dependente de outra, é razoável assumir que isso possa ser observável por estatísticas sumárias
- Como resumir esta informação em uma única grandeza numérica?

- Como definir (e mensurar!) o grau de associação entre duas variáveis aleatórias (VAs)?
- Se uma VA é dependente de outra, é razoável assumir que isso possa ser observável por estatísticas sumárias
- Como resumir esta informação em uma única grandeza numérica?

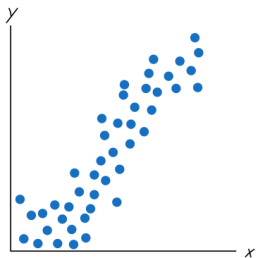
- Como definir (e mensurar!) o grau de associação entre duas variáveis aleatórias (VAs)?
- Se uma VA é dependente de outra, é razoável assumir que isso possa ser observável por estatísticas sumárias
- Como resumir esta informação em uma única grandeza numérica?

- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

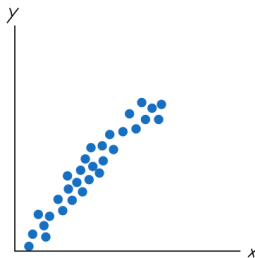
- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

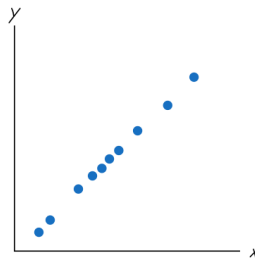
Exemplo



(a) Positive correlation
between x and y



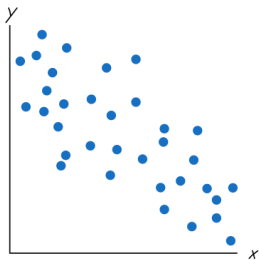
(b) Strong positive
correlation between
 x and y



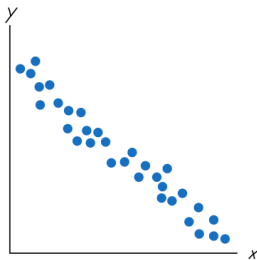
(c) Perfect positive
correlation between
 x and y

(Fonte: Triola)

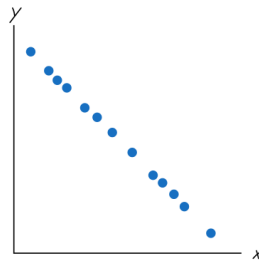
Exemplo



(d) Negative correlation
between x and y



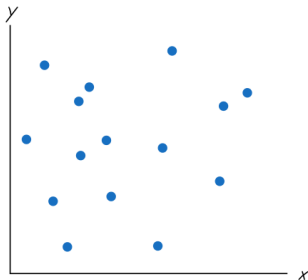
(e) Strong negative
correlation between
 x and y



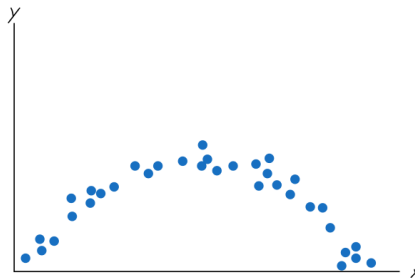
(f) Perfect negative
correlation between
 x and y

(Fonte: Triola)

Exemplo



(g) No correlation
between x and y



(h) Nonlinear relationship
between x and y

(Fonte: Triola)

- 1 Correlação
 - Associação entre duas variáveis
 - Covariância entre duas amostras
 - Coeficiente de correlação de Pearson

- Relembrando: a variância (assim como o desvio-padrão) é uma medida da dispersão da amostra
- Medida sumária que resume o quanto os dados se desviam da média
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

- Relembrando: a variância (assim como o desvio-padrão) é uma medida da dispersão da amostra
- Medida sumária que resume o quanto os dados se desviam da média
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

- Relembrando: a variância (assim como o desvio-padrão) é uma medida da dispersão da amostra
- Medida sumária que resume o quanto os dados se desviam da média
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

Definition

A covariância entre duas variáveis X e Y é uma medida de quanto ambas variam juntas (uma em relação à outra).

- Obs: duas variáveis independentes tem covariância igual a zero!

Definition

A covariância entre duas variáveis X e Y é uma medida de quanto ambas variam juntas (uma em relação à outra).

- Obs: duas variáveis independentes tem covariância igual a zero!

Definition

A correlação é a associação estatística entre duas variáveis.

Para medir essa associação, calculamos o **coeficiente de correlação r** .

- 1 Correlação
 - Associação entre duas variáveis
 - Covariância entre duas amostras
 - Coeficiente de correlação de Pearson

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1 .
- Mede a associação **linear** entre duas variáveis.
 - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1 .
- Mede a associação **linear** entre duas variáveis.
 - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1 .
- Mede a associação **linear** entre duas variáveis.
 - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1 .
- Mede a associação **linear** entre duas variáveis.
 - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

Coefficiente de correlação



Anhangüera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação

Covariância entre
duas amostras

Pearson

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

- Uma forte associação **positiva** corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação **negativa** corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A **ausência** de associação corresponde a uma correlação próxima de 0.

Example

Pesquisadores queriam entender por que a insulina varia tanto entre indivíduos. Imaginaram que a composição lipídica das células do músculo afetam a sensibilidade do músculo para a insulina. Para isto, eles injetaram insulina em 13 jovens adultos, e determinaram quanta glicose eles precisariam injetar nos sujeitos para manter o nível de glicose sanguínea constante. A quantidade de glicose injetada para manter o nível sanguíneo constante é, então, uma medida da sensibilidade à insulina.

(Fonte: Motulsky, 1995)

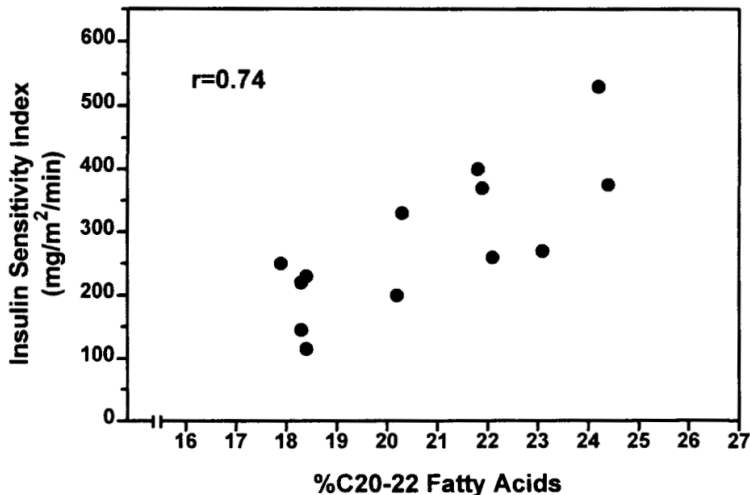
Example

Os pesquisadores fizeram uma pequena biópsia nos músculos para aferir a fração de ácidos graxos poliinsaturados que tem entre 20 e 22 carbonos (%C20-22). Como variável resposta, mediram o índice de sensibilidade à insulina.

Valores tabelados a seguir.

% C20-22 Polyunsaturated Fatty Acids	Insulin Sensitivity (mg/m ² /min)
17.9	250
18.3	220
18.3	145
18.4	115
18.4	230
20.2	200
20.3	330
21.8	400
21.9	370
22.1	260
23.1	270
24.2	530
24.4	375

Exemplo: Diagrama de dispersão dos dados



Obs: na verdade, $r = 0.77$.

- O tamanho da amostra foi $n = 13$
- Consultamos o valor crítico de r na tabela a seguir
- Testamos a H_0 que não há relação entre as variáveis na população ($H_0 : \rho = 0$).

- O tamanho da amostra foi $n = 13$
- Consultamos o valor crítico de r na tabela a seguir
- Testamos a H_0 que não há relação entre as variáveis na população ($H_0 : \rho = 0$).

- O tamanho da amostra foi $n = 13$
- Consultamos o valor crítico de r na tabela a seguir
- Testamos a H_0 que não há relação entre as variáveis na população ($H_0 : \rho = 0$).

Exemplo

TABLE A-6		Critical Values of the Pearson Correlation Coefficient r	
n	$\alpha = .05$	$\alpha = .01$	
4	.950	.999	
5	.878	.959	
6	.811	.917	
7	.754	.875	
8	.707	.834	
9	.666	.798	
10	.632	.765	
11	.602	.735	
12	.576	.708	
13	.553	.684	
14	.532	.661	
15	.514	.641	
16	.497	.623	
17	.482	.606	
18	.468	.590	

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

- Pode-se também calcular o p-valor para o coeficiente de correlação r .
- Para este exemplo, teríamos $p = 0.0021$.
- Interpretação: se não houver relação entre as variáveis (H_0), existe apenas 0.21% de chance de observarmos uma correlação tão forte com um estudo deste tamanho

- Pode-se também calcular o p-valor para o coeficiente de correlação r .
- Para este exemplo, teríamos $p = 0.0021$.
- Interpretação: se não houver relação entre as variáveis (H_0), existe apenas 0.21% de chance de observarmos uma correlação tão forte com um estudo deste tamanho

- Pode-se também calcular o p-valor para o coeficiente de correlação r .
- Para este exemplo, teríamos $p = 0.0021$.
- Interpretação: se não houver relação entre as variáveis (H_0), existe apenas 0.21% de chance de observarmos uma correlação tão forte com um estudo deste tamanho

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

- 1 o conteúdo lipídico das membranas **determina** a sensibilidade à insulina
- 2 A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
- 3 tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
- 4 as duas variáveis não são correlacionadas na população, e a estimativa observada nessa amostra é mera coincidência

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

- 1 o conteúdo lipídico das membranas determina a sensibilidade à insulina
- 2 A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
- 3 tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
- 4 as duas variáveis não são correlacionadas na população, e a estimativa observada nessa amostra é mera coincidência

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

- 1 o conteúdo lipídico das membranas determina a sensibilidade à insulina
- 2 A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
- 3 tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de **algum outro** fator (talvez algum hormônio)
- 4 as duas variáveis não são correlacionadas na população, e a estimativa observada nessa amostra é mera coincidência

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

- 1 o conteúdo lipídico das membranas determina a sensibilidade à insulina
- 2 A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
- 3 tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
- 4 as duas variáveis não são correlacionadas na população, e a estimativa observada nessa amostra é mera coincidência

- Nunca devemos ignorar a última possibilidade (erro tipo I)!
- o p-valor indica quão rara é essa coincidência
- neste caso, em apenas 0.21% dos experimentos não haveria uma correlação real, e estaríamos cometendo um erro de interpretação

- Nunca devemos ignorar a última possibilidade (erro tipo I)!
- o p-valor indica quão rara é essa coincidência
- neste caso, em apenas 0.21% dos experimentos não haveria uma correlação real, e estaríamos cometendo um erro de interpretação



- Nunca devemos ignorar a última possibilidade (erro tipo I)!
- o p-valor indica quão rara é essa coincidência
- neste caso, em apenas 0.21% dos experimentos não haveria uma correlação real, e estaríamos cometendo um erro de interpretação

- ◀ ◻ ▶ ◀ ▢ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ 🔍 ↺

- Relembrando: calculamos a variância de uma amostra para saber a dispersão dos dados
- Sua interpretação é confusa, portanto preferimos usar o desvio-padrão
- No caso do r é o contrário: a interpretação de r^2 é mais simples
- Obs: o valor r^2 também é chamado **coeficiente de determinação**, como veremos a seguir.

Elevando o r ao quadrado



Anhanguera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação
Covariância entre
duas amostras
Pearson

- Relembrando: calculamos a variância de uma amostra para saber a dispersão dos dados
- Sua interpretação é confusa, portanto preferimos usar o desvio-padrão
- No caso do r é o contrário: a interpretação de r^2 é mais simples
- Obs: o valor r^2 também é chamado **coeficiente de determinação**, como veremos a seguir.

- A set of small navigation icons typically found in Beamer presentations, including symbols for back, forward, search, and other slide controls.

- ◀ ◻ ▶ ◀ ▢ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻

Interpretando o r^2



Anhangüera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação
Covariância entre
duas amostras
Pearson

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Interpretando o r^2



Anhangüera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação
Covariância entre
duas amostras
Pearson

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Interpretando o r^2



Anhangüera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação
Covariância entre
duas amostras
Pearson

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Interpretando o r^2



Anhangüera

Medidas de
associação

Felipe
Figueiredo

Correlação

Associação
Covariância entre
duas amostras
Pearson

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!